Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey

Year: 2020 | Citations: 683 | Authors: Arun Das, P. Rad

Abstract

Nowadays, deep neural networks are widely used in mission critical systems such as healthcare, self-driving vehicles, and military which have direct impact on human lives. However, the black-box nature of deep neural networks challenges its use in mission critical applications, raising ethical and judicial concerns inducing lack of trust. Explainable Artificial Intelligence (XAI) is a field of Artificial Intelligence (AI) that promotes a set of tools, techniques, and algorithms that can generate high-quality interpretable, intuitive, human-understandable explanations of AI decisions. In addition to providing a holistic view of the current XAI landscape in deep learning, this paper provides mathematical summaries of seminal work. We start by proposing a taxonomy and categorizing the XAI techniques based on their scope of explanations, methodology behind the algorithms, and explanation level or usage which helps build trustworthy, interpretable, and self-explanatory deep learning models. We then describe the main principles used in XAI research and present the historical timeline for landmark studies in XAI from 2007 to 2020. After explaining each category of algorithms and approaches in detail, we then evaluate the explanation maps generated by eight XAI algorithms on image data, discuss the limitations of this approach, and provide potential future directions to improve XAI evaluation.